Warszawa, 01.03.2023r.

Mikołaj Markiewicz
nr. albumu doktoranta: 6132

Streszczenie rozprawy w języku polskim

Rozmiary różnych zbiorów danych gromadzonych na świecie gwałtownie rosną. Dane te są składowane w oddzielnych, niezależnych lokalizacjach. Z tego powodu wzrasta liczba nowych algorytmów do przetwarzania tak rozproszonych danych, w szczególności na potrzeby zadań klasyfikacji i grupowania. Nie ma jednak ustandaryzowanego sposobu walidacji takich algorytmów. Typowo są one testowane na niezależnych, lecz identycznie rozłożonych danych (IID), które są równomiernie rozproszone. Jednakże prawdziwy rozkład danych między niezależne węzły jest zazwyczaj nieznany, co wpływa na wyniki jak i samo przetwarzanie. Ta praca jest poświęcona poprawie oceny rozproszonych algorytmów przez zastosowanie nowych metod nierównomiernego (non-IID) partycjonowania danych. Ma to na celu złagodzenie wpływu niezdefiniowanego rozkładu danych na wyniki działania algorytmu przez ukazanie jego niedoskonałości. Ponadto w pracy zaproponowano standardowy zbiór nowych strategii partycjonowania do symulowania rozkładu non-IID na potrzeby ewaluacji algorytmów rozproszonych celem ujawnienia ukrytych niedoskonałości w ich przetwarzaniu i poprawienia jakości testowania.

Dodatkowo przedstawiona została nowa, prosta w użyciu i rozszerzalna platforma umożliwiająca dynamiczne dołączanie komponentów. W pracy zaadresowany został tym samym problem braku kompleksowych narzędzi do oceny (benchmarkingu) metod rozproszonej eksploracji danych (DDM) w rzeczywistym środowisku rozproszonym. Proponowana platforma umożliwia właściwą analizę wyników pod kątem wpływu na algorytm różnych aspektów przetwarzania takich jak końcowa jakość, obciążenie transferu i szczegółowe pomiary czasu w kolejnych etapach przetwarzania.

Poza tym w niniejszej pracy ujawniony został negatywny wpływ różnych rozkładów danych na jakość wyników przetwarzania algorytmów rozproszonych na przykładzie ewaluacji zestawu metod rozproszonej klasyfikacji i grupowania. Wprowadzono również zaadaptowany algorytm hybrydowy, który jest w stanie osiągnąć wysoką jakość niezależnie od specyficznego rozkładu danych i bez konieczności przyjmowania założeń o jednolitym rozkładzie przetwarzanych danych.

Na koniec przedstawione zostały wyniki wraz ze szczegółowym omówieniem eksperymentów, które potwierdzają zasadność badań, użyteczność proponowanego narzędzia oraz pokazują ulepszoną metodę grupowania rozproszonych danych.

……
Podpis